Assignment of ICD-10 Codes to Diagnoses in Hospital Patient Records in Bulgarian

Svetla Boytcheva

State University of Library Studies and Information Technologies
119, Tzarigradsko Shosee, 1784 Sofia, Bulgaria
svetla.boytcheva@gmail.com

Abstract. This paper briefly presents a system called EVTIMA which extracts structured data from hospital patient records in Bulgarian language. Within this context, the article considers in more detail an important extraction task: the recognition of ICD diagnose codes in the free text of patient records and their automatic assignment to text fragments. The paper summarises achievements in ICD codes recognition and presents the evaluation of current EVTIMA results.

Keywords: information extraction from hospital patient records, automatic indexing by ICD-10 codes

1 Introduction

Development of Natural Language Processing (NLP) applications in medical informatics is a hot topic nowadays. Many research prototypes and industrial systems deal with patient records (PR) texts and aim at their automatic processing. These systems use language-specific linguistic resources as well as declarative knowledge descriptions which are often system-specific too. However, the medical domain is rather complex in general; there are no standardised templates of medical documents and the existing patterns are often limited to sub-domains; in addition content formalisation is fairly complicated as it requires much background knowledge. Various natural languages and medical traditions pose specific NLP challenges. For instance, two Bulgarian-specific particularities which hamper the automatic text processing of hospital PRs are the following ones:

- The PR texts contain mixture of abbreviations and medical terminology in various languages - Bulgarian, Latin and Latin terms transliterated to Cyrillic;
- Almost no background linguistic resources in Bulgarian are available which means that the NLP modules should be elaborated from scratch together with the corresponding language and conceptual resources oriented to NLP of medical texts.

Despite all difficulties to process automatically the narrative texts in the medical domain, the interest in the development of fundamental NLP methods for medical text analysis is constantly growing. This is due to the fact that NLP is viewed as the only means to structure free-text information and provide automatic technologies for (partial) understanding of medical documents [1].

This article presents work in progress: the development of a research prototype, called EVTIMA, which extracts structured information from the free text of hospital PRs in Bulgarian language. The paper considers in more detail the (semi-)automatic procedures for recognising the terminology of the International Classification of Diseases revision 10 (ICD-10) in the PR text. The remaining part of the article is structured as follows: section 2 summarises related approaches for automatic assignment of ICD codes and section 3 briefly discusses the project background. Section 4 deals with the automatic recognition of ICD diagnose codes and presents current evaluation results. The conclusion and plans for further work are given in section 5.

2 NLP and Automatic Coding

Diseases are often described in the medical patient records as free text using terms which have to be mapped to controlled vocabularies. In this way the task of diseases recognition (which practically means e.g. assigning standardised ICD codes to diseases' names) is an important NLP challenge. The review [1] considers the automatic indexing by ICD codes as a specific NLP task that is implemented in various software systems ranging from research prototypes to commercial computer-assisted coding environments. Actually an NLP system can be designed to recognise the terminology of various medical nomenclatures and classifications but here we consider ICD since it is the only nomenclature available in Bulgarian and our task is related to the automatic assignment of ICD codes to PR diagnoses.

The nomenclature ICD or ICD CM (International Classification of Diseases with Clinical Modification), supported by WHO (World Health Organisation), is translated to many languages and serves as the main source for diagnoses definition. Subsequent ICD versions were used worldwide during the last decades: versions 8, 9 and 10. In this way we find NLP projects which approach the tasks of automatic coding using various ICD revisions. The reported performance accuracy is often presented by the so-called *F-score* (the harmonic mean of system's *Precision* and *Recall*, where *precision* is the percentage of correctly extracted entities as a subset of all extracted entities and *recall* is the percentage correctly extracted entities as a subset of all entities available in the corpus. Automatic coding has been previously approached by various systems:

- The article [2] overviews the *2007 Computational Medicine Challenge* where about 50 participants submitted results. The competition was run on anonymised radiology reports with a training set of 978 documents and a test set of 976 documents. These sets contained 45 ICD-9-CM labels; the system performance was evaluated using specific cost-sensitive *F*-metrics which incorporate penalties for over-coding (a false positive) and under-coding (a false negative). The top performing systems achieved 0,8908 accuracy, the minimum was 0,1541, and the mean was 0,7670. There were 21 systems with success between 0,81 and 0,90. Another 14 systems achieved accuracy higher than 0,70. The systems were not ranked identically by the various measures, but the differences are small in each case. As [2] emphasizes, the three top systems processed the negation, hypernyms and synonyms in some way and exploited the UMLS structure. Regarding the approach, these systems performed symbolic computations and two of them had in addition some machine-learning components. The overview notes the importance of rule-based text analysis in the coding-oriented NLP tasks.

- A recent system is MIDAS (Medical Diagnosis Assistant) [3], an advanced expert system that is able to suggest medical diagnosis esp. ICD-9-CM codes after partial analysis of radiological reports. MIDAS combines procedures for partial text understanding, called Information Extraction (IE) and machine learning from clinical histories of previously diagnosed patients.
- SynDiKATe [4] combines text parsing with semantic information derived from a Bayesian network and assigns diagnostic ICD-9-CM codes to the free text of admission diagnoses, with about 76% F-score.
- The system reported in [5] uses a hybrid approach combining example-based classification and a simple but robust classification algorithm (naive Bayes) in order to improve the efficiency of diagnostic coding. The paper [5] presents experiments run on 22 mln PRs. The reported performance results are as follows: about 48% of the medical records at Mayo clinic are automatically classified to codes with average F-score of 98,2%; another 34% of the records are classified with F-score 93,1%, and the remaining 18% of the records are classified with F-score of 58,5%. The system applies the HICDA Classification, elaborated on ICD-8 and still in use internally.
- Other systems apply *support vector machines* [6] and *Bayesian Ridge Regression* [7] for automatic associating codes for diagnosis. The article [8] compares three machine learning methods on radiological reports and points out that the best F-score is 77%.

When designing our solutions for processing PR texts in Bulgarian language, we keep in mind the lessons learned about other natural languages as well as the gains of applying various AI techniques for processing the language-independent entities extracted from the medical text. Using available modules for text analysis, our system EVTIMA performs a significant amount of symbolic computations.

3 Project Background - the EVTIMA System

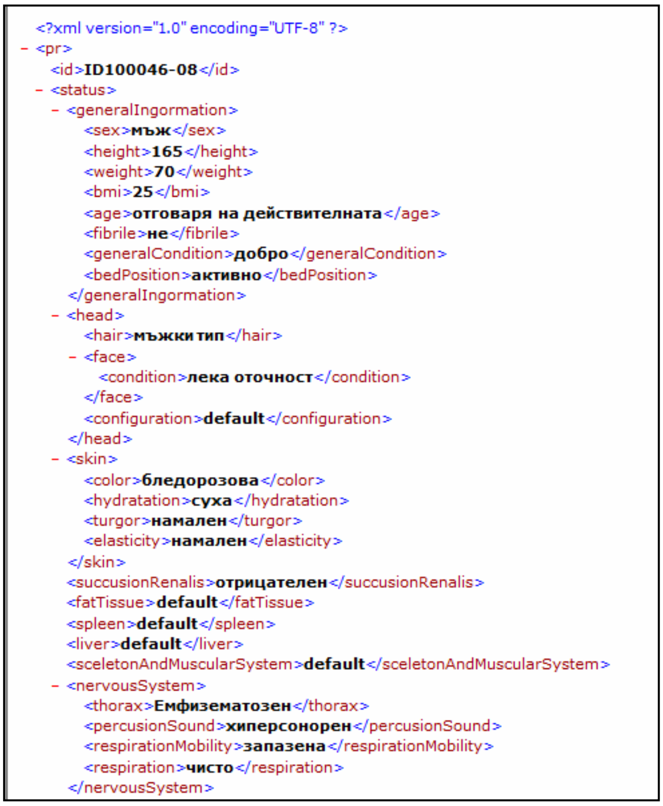
The main objective of EVTIMA project, to be achieved by 2011, is searching for complex relations in hospital PRs [9, 10]. The idea is to extract and structure entities of interest into internal templates and to process the conceptual archive by artificial intelligence methods. In order to accomplish this task we need first to process the free PR texts. EVTIMA deals with anonymised PRs provided by the Hospital Information System (HIS) of the University Specialised Hospital for Active Treatment of Endocrinology “Acad. I. Penchev” at Medical University, Sofia. The current system is based on an IE prototype which extracts important facts from the free texts of hospital PRs of diabetic patients.

The PRs texts are organised in zones: *(i)* document identification; *(ii)* personal details; *(iii)* diagnoses of the leading and accompanying diseases; *(iv)* anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors; *(v)* patient status, including results from physical examination; *(vi)* laboratory and other tests findings; *(vii)* consult with other clinicians and *(viii)* discussion (some zones might be omitted in the PRs). In this paper we consider zone *(iii)* where the diagnoses are listed as narrative text, sometimes even without punctuation marks as separators.

The prototype system works on single and multiple PRs and provides:

- exploration and update of the terminology bank;
- automatic segmentation of the PR file/s into zones;
- feature and relation extraction from the separate zones by text analysis;
- side-by-side document comparison;
- retrieval of analised documents by given features;
- export of structured representations.

Internally, the extracted information is stored in XML format (see Fig. 1) and can be exported in two different formats – they are either original texts enriched with annotation information or extracts of structured information captured during the analysis. In this way the prototype supports the development of annotated training corpus as well as the actual IE process.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <pr>
    <id>ID100046-08</id>
  - <status>
    - <generalIngormation>
        <sex>мъж</sex>
        <height>165</height>
        <weight>70</weight>
        <bmi>25</bmi>
        <age>отговаря на действителната</age>
        <fibrile>не</fibrile>
        <generalCondition>добро</generalCondition>
        <bedPosition>активно</bedPosition>
      </generalIngormation>
    - <head>
        <hair>мъжки тип</hair>
      - <face>
          <condition>лека оточност</condition>
        </face>
        <configuration>default</configuration>
      </head>
    - <skin>
        <color>бледорозова</color>
        <hydratation>суха</hydratation>
        <turgor>намален</turgor>
        <elasticity>намален</elasticity>
      </skin>
      <succusionRenalis>отрицателен</succusionRenalis>
      <fatTissue>default</fatTissue>
      <spleen>default</spleen>
      <liver>default</liver>
      <sceletonAndMuscularSystem>default</sceletonAndMuscularSystem>
    - <nervousSystem>
        <thorax>Емфизематозен</thorax>
        <percusionSound>хиперсонорен</percusionSound>
        <respirationMobility>запазена</respirationMobility>
        <respiration>чисто</respiration>
      </nervousSystem>
```

Fig. 1. Patient status data structured in XML-format extracted from PR zone *(v)*

4 Auto-Assigning ICD-10 Codes to Diagnoses

An ICD-10 code includes a letter followed by 2- to 4-digit number with a decimal point after the second digit. Main codes are organized in a hierarchy but there are also consecutive numbers after the decimal point (Table 1). The patient records in our corpus often contain diagnoses expressed as free text despite the fact that the present Hospital Information System offers menu choice for ICD diagnoses; actually the past diseases are still recorded in the personal history as free text. Thus in the PR text we find terms, phrases and paraphrases which differ significantly from the ICD disease description. The number of diagnoses per patient varies from 1 to about 20 (see examples 1-2). Thus the automatic association of ICD codes to diagnoses is a challenging effort. Below we discuss excerpts of PR texts.

(E10 - E14)	Diabetes mellitus
E10	Insulin-dependent diabetes mellitus
E10.0	With coma
E10.1	With ketoacidosis
E10.2	With renal complications
	......
E11	Non-insulin-dependent diabetes mellitus

Table 1. Extract of ICD-10 Codes for *diabetes mellitus*

Example 1:

Тиреоидит на Хашимото - хипортиреоидна фаза, медикаментозно компенсиран.

Example 2:

ДИАГНОЗА: Захарен диабет тип 1. Диабетна дистална симетрична полиневропатия. Сърдечна автономна невропатия. Диабетна пролиферативна ретинопатия. Състояние след лазеркоагулация. Състояние след операция по повод двустранна катаракта. Начална диабетна нефропатия. Диабетна макроангиопатия. Периферна съдова болест. ИБС-стабилна стенокардия. Мозъчно-съдова болест. Дислипидемия. Тиреоидит на Хашимото, еутиреоидна фаза. Витилиго. Болест на Адисон. Генерализирана остеопороза. Състояние след фрактура на дясна бедрена шийка и дясна предмишница. Секторална резекция на дясна млечна жлеза за фиброаденом.

Table 2 shows that generally defined diagnoses in the PR texts can be associated with several ICD codes (like *hypothyroidism*), in contrast to diagnoses which can be matched exactly to ICD disease (for instance, *diabetic polyneuropathy*). It is well known that in cases of ambiguity or unclear/partial/too general disease description, human annotators can assign different codes to the same text. The best NLP practices are based on gold standards and precisely annotated training corpora but we are at the very beginning of our IE coder development and present here only our initial results.

Diagnose	Possible ICD-10 Codes	
Hypo-thyroidism	E02	Subclinical iodine-deficiency hypothyroidism
	E03	Other hypothyroidism
	E03.0	Congenital hypothyroidism with diffuse goitre
	E03.1	Congenital hypothyroidism without goitre
	E03.2	Hypothyroidism due to medicaments and other exogen
	E03.3	Postinfectious hypothyroidism
	E03.8	Other specified hypothyroidism
	E03.9	Hypothyroidism, unspecified
	E05	Thyrotoxicosis [hyperthyroidism]
	E89.0	Postprocedural hypothyroidism
Diabetic polyneuropathy	G63.2	Diabetic polyneuropathy

Table 2. Lexical correspondences between PR diagnoses and ICD-10 terms

Sometimes there are no common words at all between the phrasal disease description in the PR text and the terminology used in ICD-10 codes. Let us consider the following example:

Example 3:

Diagnosis:

Състояние след байпас на подбедриците двустранно.

(Post bypass surgery symptoms of shanks bilateral)

The closest ICD code that can be associated with this diagnose is:

I97.89 Постпроцедурни болести на органите на кръвообращението, некласифицирани другаде

I97.89 Other postprocedural complications and disorders of the circulatory system, not elsewhere classified

However, there are no common words between the diagnosis and the description of ICD code I97.89, therefore matching the PR phrase to the ICD disease names is useless in this case.

There are also diagnoses described by Latin terms or by Latin terms transliterated to Cyrillic letters, for instance:

Статус пост инфарктус миокардий.
ХИПОТИРЕОИДИЗМУС АУТОИМУНЕС.
ДИАБЕТЕС МЕЛИТУС ТИПУС 2.

One also finds in the PR text many abbreviations of Bulgarian and Latin terminology for diagnoses.

The main obstacle for the automatic assignment of ICD codes is that the correspondence between the PR diagnoses and ICD-10 names/codes is not limited to 1:1 (one to one) relation. There are many cases of 1:*n* (one to many), *n*:1 (many to one) and *n*:*m* (many to many) relations. Although the language used in the PRs is generally restricted and the diagnoses are explained by specific terminology, there is a variety of linguistic expressions in the free PRs text which might refer to the same diagnose. EVTIMA uses a 3-steps pipeline algorithm for automatic coding of diagnoses (see Table 3).

Step 1: *Shallow text analysis by Regular expressions and Patterns matching*

Step 2: *Searching diseases names in the terminology resource bank:*

- medical terminology dictionary
- abbreviations rules
- Latin – Cyrillic transliteration rules

Step 3: *Application of manually added terminology binding rules by experts*

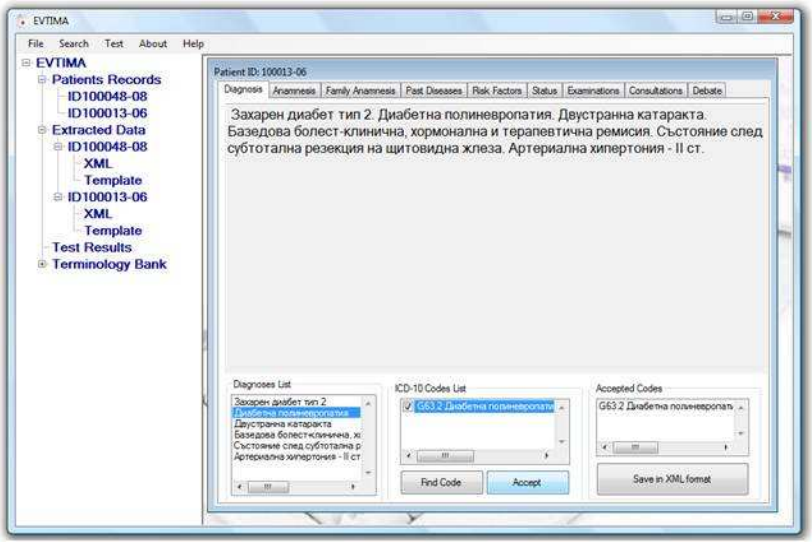
Table 3. Algorithm for coding diagnosis

At step 1 our algorithm applies regular expressions (RE) in order to associate a PRs diagnose with ICD-10 code. The matching procedure skips all blank spaces, it is not case-sensitive. There are few preferences associated to all REs. Higher priority is assigned to the REs, which enable exact matches of all the words found in the PR diagnose and the name, associated with the ICD code, contains no additional words. For more specific or more general explanations the REs match either subset of the PR diagnose words or the code specification includes additional words, not presented in the PR text. This match is applied like a greedy algorithm. In case that the RE-based analysis fails, another matching procedure starts which searches for word derivatives using simple derivation rules for Bulgarian language. There are also some rules that deal with paraphrasing, they cope with different word order or ‘weaken’ the diagnose explanation (by skipping some words) and try to match it to more general ICD-10 diagnose.

At step 2 the algorithm searches the system terminology resource bank only for diagnoses expressed in some non-standard manner. The constantly growing resource bank includes at present:

- A medical terminology dictionary – terms with associated relations like *synonym, equivalent, general, specific*,
- Some rules for abbreviations of the terms met in the PR texts,
- Latin – Cyrillic transliteration rules, and
- Latin – Bulgarian medical terminology dictionary.

A taxonomy of diabetes is adopted for our purposes using the sources at the Bioportal cite [11]. There is a lack of language and conceptual resources in this domain and our dictionaries are manually prepared for solving the current task.



Patient ID: 100013-06

Захарен диабет тип 2. Диабетна полиневропатия. Двустранна катаракта. Базедова болест-клинична, хормонална и терапевтична ремисия. Състояние след субтотална резекция на щитовидна жлеза. Артериална хипертония - II ст.

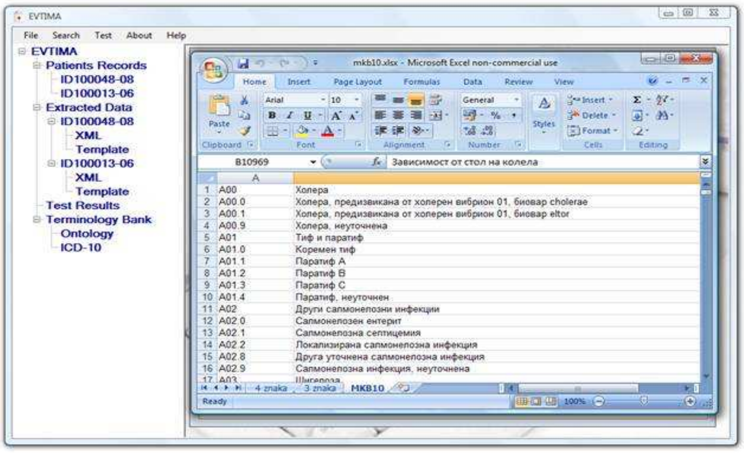
Diagnoses List:
Захарен диабет тип 2
Диабетна полиневропатия
Двустранна катаракта
Базедова болест-клинична, х
Състояние след субтотална р
Артериална хипертония - II ст

ICD-10 Codes List:
☑ G63.2 Диабетна полиневропати

Accepted Codes:
G63.2 Диабетна полиневропать

Fig. 1. EVTIMA interface – Manual assignment of ICD codes to a PR diagnose

In case that the first and second steps of the algorithm are not successful, at the third step EVTIMA considers rules for manual assignment of ICD-10 codes to diagnoses (see Fig. 1). Manually-established associations < *phrase* - *ICD code*> are stored in the system resource bank with highest priority for further code assignments. The system has two modes – automatic and manual codes association. In the automatic mode, the association algorithm runs and returns the codes suggested with highest accuracy. For



B10969 | Зависимост от стол на колела

	A	
1	A00	Холера
2	A00.0	Холера, предизвикана от холерен вибрион 01, биовар cholerae
3	A00.1	Холера, предизвикана от холерен вибрион 01, биовар eltor
4	A00.9	Холера, неуточнена
5	A01	Тиф и паратиф
6	A01.0	Коремен тиф
7	A01.1	Паратиф A
8	A01.2	Паратиф B
9	A01.3	Паратиф C
10	A01.4	Паратиф, неуточнен
11	A02	Други салмонелозни инфекции
12	A02.0	Салмонелозен ентерит
13	A02.1	Салмонелозна септицемия
14	A02.2	Локализирана салмонелозна инфекция
15	A02.8	Друга уточнена салмонелозна инфекция
16	A02.9	Салмонелозна инфекция, неуточнена
17	A03	Шигелоза

Fig. 2. EVTIMA System – Access to ICD-10 codes

the training corpus the expert can switch to manual mode for associating codes with diagnoses in order to correct and improve the rules for better system performance in further tasks. The manual mode supports also the development of the training corpus.

Sometimes the set of automatically-assigned codes is empty and the system alerts the user about this. EVTIMA provides direct access to the ICD-10 codes table (see Fig. 2) and allows to authorised users to bind one or more diagnoses to one or more codes.

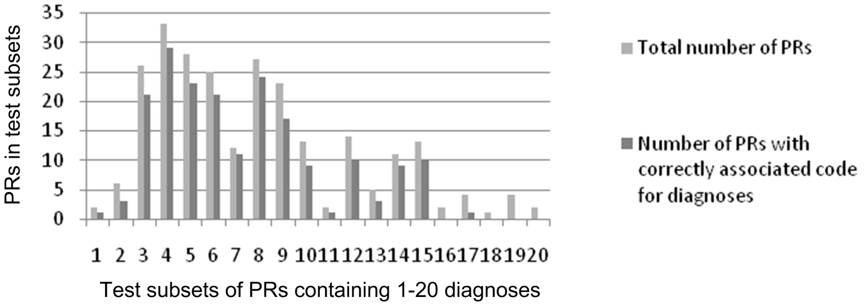


Fig. 3. Percentage of PRs with correctly associated ICD-10 codes

The algorithm for automatic assignment of ICD-10 codes for diabetic patients was trained on a corpus of 197 PRs. The test corpus contains approximately 250 unknown PRs. About 50 ICD diagnoses, related to diabetes and its complications as well as various accompanying diseases, are found in the test corpus. The whole test corpus was split into PR subsets grouped according to the number of diagnoses available in the PR text, from 1 to 20 diagnoses. This division allows for more precise evaluation because multiple diagnoses are often expressed by complex noun phrases without clear separators, which are hard to analyse and index by ICD codes. The evaluation is done separately for each test corpus subset because we have assumed that the automatic assignment is more successful when the PR text contains less diseases names.

The evaluation results are illustrated at Fig. 3 and Fig. 4. The *x*-axis of both diagrams represents the twenty PR "families" consisting of PRs with 1-20 diseases. Fig. 3 summarizes the results from the PR perspective. For some PRs, part of the diagnoses are correctly encoded and others are wrong, so Fig. 3 shows the ratio of PRs with fully associated diagnoses vs total number of PRs tested. We see that our algorithm is most accurate for the 7th family of 12 PRs, which contain 7 diagnoses each; for 11 PRs the diagnoses are assigned successfully. The worst performance is seen for the families of PRs with 16-20 diagnoses; from 16 PRs belonging to those 5 subsets, only in one PR we have correct coding which is due to the accurate punctuation marks.

The algorithm performance can be also evaluated from the perspective of recognised individual diagnoses. Fig. 4 presents the ratio of correctly associated codes for diagnoses compared to the total number of diagnosis included in the corresponding PR set.

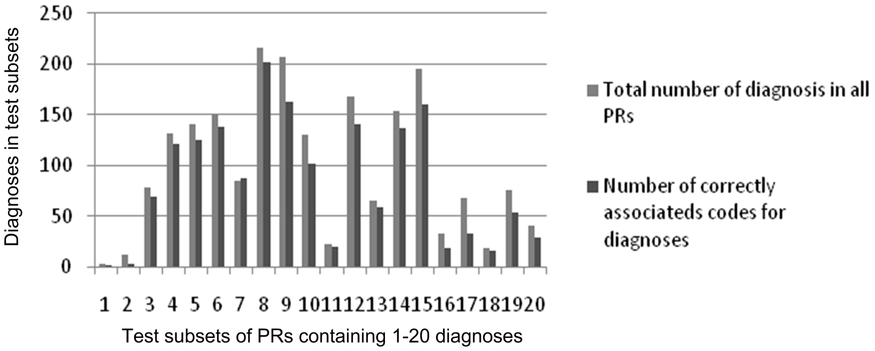


Fig. 4. Percentage of diagnoses with correctly associated ICD-10 codes

5 Conclusion and Further Work

The paper presents a software environment called EVTIMA which supports the automatic extraction and structuring of patient data from hospital PR texts. So far we have evaluated its performance on several IE tasks dealing with the recognition of diagnosis, anamnesis, patient' status and treatment zones of the PRs. The results are promising and show that partial analysis, IE and machine learning techniques are successfull approaches and need to be developed further. EVTIMA serves as well as a tool for semi-automatic annotation of the recognised features. It supports different output formats which facilitate the further processing of the obtained analysed structures. Last but not least, the environment provides a large number of functionalities for manual expert activity. This helps to correct and upgrades the system resources and to elaborate the annotation of the training corpus.

As future work we plan to develop algorithms for discovering more complex relations and other dependences among patient-related facts. Further exploration of the negative expressions and their semantic interpretation in the internal representation is also one in our future tasks. Medical patient records contain complex temporal structures which are connected to patient's case history and the family history. Another interesting question is related to the automatic recognition of illness duration, the drug admission and the related clinical patient tests.

Acknowledgements. The research work presented in this paper is supported by grant DO 02-292 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2011.

References

1. Demner-Fushman, D., W. Chapman and C. McDonald. *What can natural language processing do for clinical decision support?* Journal of Biomedical Informatics, 42(5), October 2009, pp. 760-772.
2. Pestian J, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K. B. Cohen, and D. Wlodzislaw. *A shared task involving multi-label classification of clinical free text*. In: ACL'07 workshop on biological, translational, and clinical language processing (BioNLP'07). Prague, Czech Republic; 2007, p. 36–40.
3. Sotelsek-Margalef, A. and J. Villena-Román. *MIDAS: An Information-Extraction Approach to Medical Text Classification (MIDAS: Un enfoque de extracción de información para la clasificación de texto médico*), Procesamiento del lenguaje Natural n. 41, 2008, pp. 97-104.
4. Hahn, U., K. Schnattinger and K. Markó. *Wissensbasiertes Text-Mining mit SynDiKATe (Knowledge based text mining with SynDiKATe)*. Künstliche Intelligenz, vol. 2, 2002.
5. Pakhomov, S., J. Buntrock and C. G. Chute. *Automating the assignment of diagnosis codes to patient encounters*, Journal of American Medical Informatics Association, 13, 2006, pp. 516-52.
6. Farkas, R. and G. Szarvas. *Automatic Construction of Rule-based ICD-9-CM Coding Systems.* In: C. Baker and Su Jian (Eds.), Proceedings of the Second International Symposium on Languages in Biology and Medicine (LBM) 2007, BMC Bioinformatics, 2008, 9 (Suppl. 3), available online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2352868/.
7. Crammer, K., M. Dredze, K. Ganchev, and P. Talukdar. *Automatic Code Assignment to Medical Text*. In: Proceedings of BioNLP Workshop at the conference of the Association for Computational Linguistics ACL-2007.
8. Coffman, A. and N. Wharton. *Clinical Natural Language Processing: Auto-Assigning ICD-9 Codes*. Overview of the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge. Available online at http://courses.ischool.berkeley.edu/i256/f09/Final%20Projects%20write-ups/coffman_wharton_project_final.pdf
9. Boytcheva, S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova. *Extraction and Exploration of Correlations in Patient Status Data*. In: Biomedical Information Extraction, In: Proceedings of the International Workshop held in conjunction with RANLP-09. Borovets, Bulgaria, September 2009, pp. 1-7.
10. Boytcheva, S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova. *Structuring of Status Descriptions in Hospital Patient Records.* In the Proc. 2nd International Workshop on Building and Evaluating Resources for BioMedical Text Mining, associated to the 7th International Conference on Language Resources and Evaluation (LREC-2010), Malta, May 2010, pp. 31-16.
11. BioPortal, http://bioportal.bioontology.org/visualize/13578/Diabetes_Mellitus